Volume 15 Number 1 1987 Nucleic Acids Research

Unusual features of transcribed and translated regions of the histone H4 gene family of *Tetrahymena thermophila*

Stuart Horowitz[1], Josephine K.Bowen, Gary A.Bannon[2] and Martin A.Gorovsky*

Department of Biology, University of Rochester, Rochester, NY 14627, USA

Received August 18, 1986; Revised October 27, 1986; Accepted November 17, 1986

ABSTRACT

The complete DNA sequence is presented of H4-II, the second of the pair of histone H4 genes of the ciliated protozoan, Tetrahymena thermophila. Both H4 genes code for the same protein. Codon usage in these and other Tetrahymena genes is severely restricted and is similar to that in yeast. Flanking regions are AT-rich (>75%), relative to coding sequences (~45% GC). Except for small, similarly positioned homologies, flanking sequences of the two genes are different. Canonical sequences in higher eukaryotic promoters are not obvious in these genes. Instead, short, localized, base composition eccentricities characterize the 5' flanking sequences of all Tetrahymena genes analyzed. The consensus, $PyPu(A)_{3-4}$ ATGG initiates translation in these and all other known Tetrahymena genes. Nuclear transcripts and messages of both growing and starved cells begin at multiple sites, mainly at the first or second A residue following a pyrimidine. The palindrome typical of histone message 3' termini in higher organisms is not present. Downstream of both genes are sequences similar to the processing/polyadenylation signal of higher eukaryotes, although the unique 3' ends are not those predicted by the location of the signals.

INTRODUCTION

All of the core histone proteins of the ciliated protozoan Tetrahymena have been sequenced (1-4); their primary structures are the most divergent known (5). Recent studies have also demonstrated that ciliate nuclear genes (including histone genes of Tetrahymena) have a unique genetic code in which the "universal" stop codon, TAA, codes for glutamine (6-10). These observations suggest that ciliates diverged from other eukaryotes very early, possibly at a time when some of the fundamental features of eukaryotic gene organization and expression were still being elaborated. Thus, a detailed analysis of one or more conserved genes or multigene families in ciliates should shed light on the evolution of eukaryotic genes, and may reveal new mechanisms of gene organization and regulation. Alternatively, knowledge of the evolutionarily primitive state of fundamental elements of eukaryotic genes could provide insights into their function in more complex eukaryotes.

Although ciliate genes transcribed by RNA polymerases I or III have been characterized in some detail (11,12), surprisingly little is known about ciliate genes transcribed by pol II. To our knowledge, the complete gene sequences of only three different genes in three different genera have been reported: Oxytricha actin (13), Tetrahymena histone H4 (14) and Stylonychia α-tubulin (7). Partial sequences of four other genes, two for non-allelic Paramecium surface antigens (6,10) and for two Tetrahymena non-allelic H3 genes (8) have also been described.

In an attempt to characterize at least one ciliate multi-gene super-family in detail, we have initiated studies on the histone genes in the most intensively studied ciliate species, Tetrahymena thermophila. Histone genes offer a number of advantages for such studies. 1) The histone multigene super-family is ubiquitous in eukaryotes and has been intensively analyzed in a variety of other species. As a result, heterologous probes and a large amount of comparative data are immediately available. 2) More protein primary sequence data exist for histones than for any other ciliate gene family (1-4), allowing unambiguous assignment of codons (8). 3) Detailed two dimensional gel analyses (5) make it likely that most, if not all, of the primary sequence variants have been identified for all of the histone types and that the histone gene super-family consists of a sizeable but not unmanageable number of about 10-12 members having distinct primary sequences. 4) Initial studies indicated that the number of genes for H3 and H4 histones in Tetrahymena was only slightly larger than the number of primary sequence types (15), offering the possibility of complete analysis. 5) Initial studies also indicated that these genes were differentially regulated in a manner similar to that in mammalian cells (15).

The first published sequence of a Tetrahymena pol II gene, H4-I, corresponds exactly to the major H4 protein of Tetrahymena (14,3). Interestingly, the only case in which more than one primary sequence variant of histone H4 within an organism has been reported is in Tetrahymena pyriformis (3). It was of particular interest, therefore, to clone and sequence H4-II, the second of the pair of genes (14) that makes up the histone H4 gene family in T. thermophila, to determine whether this heterogeneity was conserved (in this closely-related species) and therefore likely to be functional, or was due to chance fixation of a presumably neutral mutation.

In this study we present the complete DNA sequence of H4-II and its immediately flanking regions. These sequences are compared to those of H4-I, to other ciliate genes and to histone genes of other organisms. We also

present the first data mapping the ends of messages and of nuclear RNA transcripts of a ciliate gene.

MATERIALS AND METHODS

Cells and Culture Conditions

Tetrahymena thermophila were grown axenically in enriched proteose peptone at 28° C as described (16). Log-phase cells were grown to a density of ~2 x 10^5 cells/ml. Cells were starved by washing log-phase cells twice and resuspending them at concentrations of ~2 x 10^5 cells/ml in 10mM Tris-Cl (pH = 7.4). Cells were starved at 28° for 20-24 hrs.

DNA Isolation

Macronuclei were isolated from log-phase cells essentially as described in Gorovsky et al (16). DNA was isolated as described by Bannon et al. (14) except that Proteinase K (Merck) was used in place of Pronase and incubation was at 65°C.

Cloning of Tetrahymena Histone Genes

Histone H4 gene-containing clones were isolated from a recombinant genomic library of *Tetrahymena* macronuclear DNA inserted into phage λ gt WES·λb, constructed as previously described (8). Of 10 H4 gene-containing clones, 8 were identical to a previously isolated H4 gene, H4-I (14). Two clones contained the second H4 gene, H4-II and a closely linked H3 gene, H3-II (8). Recombinant phage DNA was isolated for further characterization and subcloning as previously described (8).

DNA Sequencing

Recombinant M13 phages were generated for sequencing as described previously (8). Additionally, data gained from initial sequencing were used to develop clones for sequencing the 3' flanking DNA of H4-II. Recombinant M13 phages were sequenced by the dideoxy method (17) using the BRL sequencing kit, according the the manufacturer's instructions. DNA sequence data presented were checked by sequencing overlapping subclones, sequencing both strands of DNA or sequencing at least two independent subclones of the same fragment (except for about 350 bp at the 3' terminus of H4-II, where one clone was sequenced twice).

Computer Analyses of DNA Sequences

DNA sequence data were entered directly into a Xerox 820-II microcomputer via acoustic digitizer. Gels were merged, edited, and all sequence homology searches were performed using Pustell-Kafatos sequence analysis software (18). Specific parameters of the searches are described in the figure legends where

appropriate. DNA sequences for other H4 genes were found in the Genbank and EMBL databases using the Bionet computer resource (IntelliGenetics). Histone H4 genes were transferred from Bionet to an IBM PC and edited to remove all non-coding sequences and termination codons. H4 genes were then linked in-phase and translated using Pustell-Kafatos software to generate a codon usage table.

RNA Mapping of H4-I

Tetrahymena RNA was isolated according to the methods of Calzone et al (19). Poly A^+ RNA was prepared by affinity chromatography on oligo dT cellulose, poly U Sepharose (20), or Amersham Hybond-mAP (according to instructions).

RNase protection mapping was performed as described by Melton et al. (21) and the Riboprobe system protocols provided by Promega Biotec. Anti-sense ^{32}P-labeled RNA transcripts were prepared using the Amersham SP 6 system with plasmids containing the appropriate flanking regions of H4-I or H4-II (see figure legends). Cold UTP was included in the reactions and transcripts were run on acrylamide gels, dried and autoradiographed to assay production of full-length transcripts. Optimum RNase digestion conditions were empirically determined to be 4 µg RNase A/ml and 90 units RNase T1/ml at 0° for 10 min.

Primer extension using reverse transcriptase of avian myeloblastosis virus followed the procedures of Ghosh et al. (22). All primers were ^{32}P-5'-end-labeled with T4 polynucleotide kinase and [^{32}P]-γ-ATP. Two synthetic oligomers were used as primers; both were annealed in 0.1 M NaCl, 20 mM Tris-HCl (pH = 7.9), 0.1 mM EDTA, at 37° for 4 hrs. Primer 1 corresponds to the H4-I sequence from position +6 to -14 (Fig. 1); primer 2 from +47 to +28. The third primer was derived from a 91 bp Hae III to Dde I restriction fragment, (position +5 to +95 of H4-I), which was strand-separated and annealed in 0.3 M NaCl, 40 mM PIPES (pH = 6.4), 1 mM EDTA, 25% formamide at 53°C for 16 hrs.

S1 nuclease mapping was performed with a Dde I fragment of H4-I (position -168 to +95) as described by Benyajati and Dray (23). Hybridization conditions were 0.3 M NaCl, 40 mM PIPES (pH = 6.4), 1 mM EDTA, 25% formamide at 44°C for 16 hr.

RESULTS AND DISCUSSION

The Coding Region

The sequence of H4-II and its flanking regions is shown in Fig. 1; the previously published sequence of H4-I is also included for comparison. The derived amino acid sequence of both genes is identical and matches that of the

major of the two H4 proteins of the related species, Tetrahymena pyriformis. Since there are only two H4 genes in Tetrahymena thermophila (14), there can be only one H4 protein. It is likely therefore, that the reported heterogeneity of the H4 protein in T. pyriformis represents fixation of a neutral mutation rather than functionally significant variation, and (except for this case) H4 remains unique among the histones in its lack of variation between cell types or between different physiological or developmental states within a species. The questions of why and how Tetrahymena independently regulates two H4 genes (15) coding for the same protein remains an intriguing one.

Translation of H4-II initiates with ATG and terminates with TGA, as it does in all other sequenced ciliate genes. The DNA sequences of the coding regions of H4-I and H4-II are remarkably similar, differing at only 14 of 306 bp, suggesting the genes are either recently duplicated or have been recently corrected against each other. The nucleotide substitutions between the two genes do not appear to be distributed randomly (Fig. 1). Similar clustering of differences between two non-allelic H3 genes (8) is also discernible. Correction by gene conversion over short patches of DNA has recently been described (24) and might account for the pattern of homology seen here. For both H4-I and H4-II (see below) and for two H3 genes (8), sequence homology is much lower in the flanking regions than in the coding regions, arguing against the notion that recent duplication is responsible for the similarities in nucleotide sequence between non-allelic histone genes in Tetrahymena. An RNA-mediated mechanism that might act to maintain strong homology among dispersed members of a gene family has recently been suggested by Doolittle, based on considerations of unlinked 5S genes in Neurospora (25). In this respect, it should be noted that H4-I and H4-II are on different chromosomes (14). However, we do not think that Tetrahymena H4 genes are corrected by this mechanism, since homology does not extend to transcribed regions which are untranslated (Fig. 1). A possibility other than duplication or correction is that H4-I and H4-II coding sequences are functionally constrained from divergence, such as by tRNA availability (see below).

Codon usage in H4-I and H4-II is highly biased. Of 60 possible codons which could encode the 18 different amino acids of the H4 protein, only 32 are used (Table I). A strikingly similar pattern of codon usage is seen in the two H4 genes of the yeast, Saccharomyces cerevisiae, which uses only 30 codons; 22 unused codons are shared between the two organisms. Similar codon usage explains the observation that the overall nucleotide sequence homology between the H4 genes of Tetrahymena and of yeast is the same as the protein

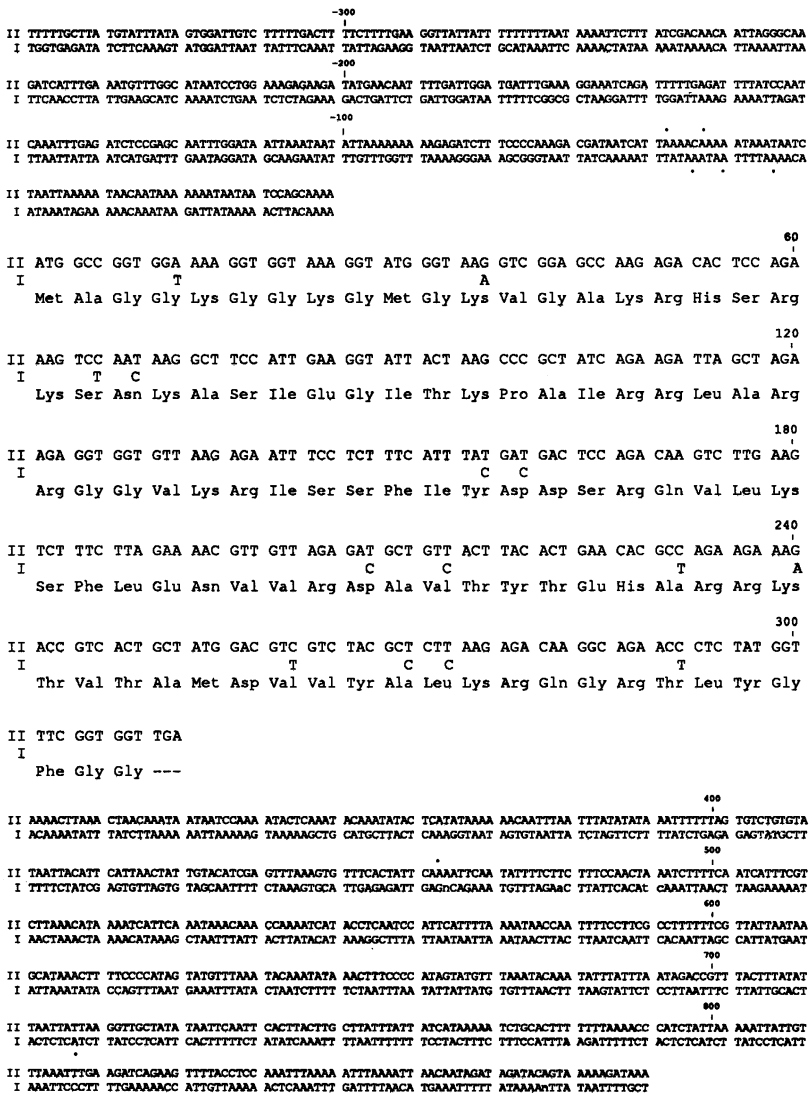

Figure 1: The DNA Sequence of Tetrahymena H4-II and H4-I. II=H4-II, I=H4-I. Only the coding strands are shown. Upstream sequences extend to -340 (ATG=1), which is the first position of translation of a divergently transcribed H3 gene (H3-II) linked to the H4-II gene. Nucleotides in the coding region are identical in H4-II and H4-I, except where indicated. Dots in the upstream and downstream regions correspond to strong sites of transcription initiation and polyadenylation/cleavage respectively, of the genes (see Figs. 2 and 8). The DNA sequence of H4-I has been published previously (14).

Table I - Comparison of Codon Usage in H4 and Other Genes.

	T	Y	C	S	A		T	Y	C	S	A		T	Y	C	S	A		T	Y	C	S	A
TTT Phe	0	0	2	95	0	TCT Ser	5	6	47	180	22	TAT Tyr	3	3	28	63	11	TGT Cys	0	0	56	31	1
TTC Phe	6	4	69	112	28	TCC Ser	9	6	42	126	18	TAC Tyr	5	5	50	103	41	TGC Cys	0	0	47	8	1
TTA Leu	4	4	30	92	9	TCA Ser	0	0	59	50	1	TAA ---	---	---	---	---	---	TGA ---	---	---	---	---	---
TTG Leu	2	10	28	263	18	TCG Ser	0	0	1	18	2	TAG ---	---	---	---	---	---	TGG Trp	0	0	20	39	0
CTT Leu	1	0	35	32	4	CCT Pro	0	0	24	45	5	CAT His	0	0	10	40	3	CGT Arg	0	4	0	29	23
CTC Leu	3	0	54	12	37	CCC Pro	2	0	35	18	5	CAC His	4	4	34	74	23	CGC Arg	0	0	2	4	61
CTA Leu	0	4	6	47	10	CCA Pro	0	2	35	136	2	CAA Gln	4	4	38	142	10	CGA Arg	0	0	0	7	4
CTG Leu	0	0	1	28	20	CCG Pro	0	0	2	2	26	CAG Gln	0	0	3	45	16	CGG Arg	0	0	0	3	18
ATT Ile	8	6	54	161	22	ACT Thr	9	9	97	0	15	AAT Asn	1	0	43	76	7	AGT Ser	0	0	16	32	0
ATC Ile	2	8	60	130	57	ACC Thr	3	3	67	104	51	AAC Asn	3	2	53	158	16	AGC Ser	0	0	20	16	1
ATA Ile	0	0	4	36	0	ACA Thr	0	0	66	55	5	AAA Lys	6	8	48	180	31	AGA Arg	26	23	117	225	58
ATG Met	6	2	46	111	26	ACG Thr	0	0	0	18	3	AAG Lys	16	14	145	301	110	AGG Arg	0	1	1	28	15
GTT Val	8	8	73	177	26	GCT Ala	1	7	123	302	37	GAT Asp	2	4	91	184	17	GGT Gly	23	32	112	329	79
GTC Val	10	8	67	139	56	GCC Ala	6	5	72	120	45	GAC Asp	6	2	34	148	24	GGC Gly	2	0	7	25	71
GTA Val	0	0	12	35	1	GCA Ala	0	0	31	49	7	GAA Glu	6	8	88	314	18	GGA Gly	3	0	48	25	44
GTG Val	0	0	0	32	30	GCG Ala	0	0	1	16	6	GAG Glu	0	0	35	69	32	GGG Gly	0	0	0	27	16

T = Tetrahymena H4 genes, Y = yeast (Saccharomyces cerevisiae) H4 genes (35), C = all ciliate genes sequenced: Two T. thermophila H4 genes, two partial H3 genes (8), actin (Claire Cupples, PhD Thesis, York University, Ontario, Canada) and H1 (M. Wu and M. Gorovsky, unpublished), Paramecium G and H surfce antigen cDNA's (6,10) Stylonychia alpha tubulin (7) and Oxytricha actin (13). The unusual codons TAA and TAG are not included here because there are no unusual codons in the H4 genes. S = 14 S. cerevisiae genes. The genes included in these analyses are: G3PDH, enolase, actin, H2A, H2B, TRP 5, CYC 1 and CYC 7 (26,27), H4 and H3 (35), RAD 2 (44), and RAD 3 (45) are also included here. (The RAD genes encode relatively rare mRNA's). A = all histone H4 genes sequences available: The primary data were taken from the Bionet computer resource. Only complete genes found on the Genbank or EMBO databases are included here. The organisms whose H4 genes were used are: Tetrahymena thermophila (2 genes), Saccharomyces cerevisiae (2 genes), Xenopus laevis, Xenopus borealis, Physarum polycephalum, Triticum aestivum, Mus musculus, Homo sapiens, Lytechinus pictus, Strongylocentrotus putpuratus, Gallus domesticus (13 genes total).

homology (about 75%) (14). Since codon preferences are strongly correlated with abundant isoaccepting tRNA's in yeast and in E. coli (26,27), similarities in codon usage between homologous genes in different organisms could reflect similarities in tRNA availability or the need to maintain extensive structural homology in the mRNA's (28), or both. The relative abundance of isoaccepting tRNA's should be reflected in codon usage tables for other abundant mRNAs. Codon usage in Tetrahymena H3 genes (8), Tetrahymena actin (Claire Cupples, PhD thesis, York University), a Tetrahymena H1 gene (M. Wu and M. Gorovsky, unpublished) and for part of a Tetrahymena tubulin gene, (unpublished) as well as for other relatively abundant ciliate mRNAs (6,7,13,10), are very similar to those for histone H4 (Table I), strongly suggesting that tRNA abundance plays a major role in the restricted codon usage of Tetrahymena H4 genes. Similar analyses of yeast genes (Table I) suggests that both organisms utilize overlapping sets of abundant isoaccepting tRNA's.

TABLE 2. A Tetrahymena-Specific Translation Initiation Sequence

Gene	Sequences	Ref.
TETRAHYMENA HISTONE GENES		
H4-I	ACAAAAATGG	1
H4-II	GCAAAAATGG	
H3-I	ATAAAAATGG	23
H3-II	GCAAAAATGG	23
H1[1]	ATAAAATGG	
CONSENSUS	$\frac{AC}{GT}(A)_{3-4}$ATGG	
OTHER CILIATE GENES		
TETRAHYMENA ACTIN[2]	AGTAAAATGG	
OXYTRICHA ACTIN	GTACATATGG	26
STYLONYCHIA TUBULIN	TTCATCATGG	20
OTHER CONSENSUS SEQUENCES		
OTHER HISTONE GENES	NNCANNATGG	22
NON-HISTONE GENES	NCC$\frac{A}{G}$CCATGG	27

[1] M. Wu, D. Allis, R. Richman, R. Cook, M. Gorovsky, unpublished observations.
[2] Claire Cupples, PhD Thesis. York University, Ontario, Canada.

Analysis of Table I reveals that there are two kinds of exceptions to the correlation between non-random use of degenerate codons in Tetrahymena H4 genes and in all ciliate genes; both are reflected in the use of codons found in other H4 genes sequenced. The first is the exclusion of three codons from all H4 genes: TTT (Phe), ATA (Ile) and AGT (Ser). Although the functional significance of their absence is unclear, it is unusual, considering the number of H4 genes analyzed. A similar analysis of H3 genes (data not shown), indicates that all codons are used (at least once) for the amino acids in H3 proteins. The second exception is reflected in the codons TCA (Ser), ACA (Thr) and GCA (Ala). Although these codons are acceptable to Tetrahymena's translational machinery (i.e. they are used frequently in other ciliate genes), their absence in H4-I and H4-II is correlated with their relatively infrequent use in other H4 genes. Presumably, these are examples of message-specific structural features affecting the evolution of these sequences. Taken together, we think that codon use in Tetrahymena's H4 genes is largely, but not exclusively, a reflection of tRNA availability, and that there appear to be H4-specific considerations which also play a role.

A Tetrahymena-specific translation initiation sequence

Abutting the initiation codon in both H4 genes is the sequence PuCAAAAATGG, a sequence also found in Tetrahymena H3 genes (8). A slight variation of this sequence is found in all Tetrahymena genes sequenced to date, generating the consensus sequence, $\frac{AC}{GT}(A)_{3-4}$ ATGG for the translation initiation region of Tetrahymena genes (Table II). This sequence is not found in other ciliate genes and does not fit the consensus generated from other eukaryotic non-histone (29) or histone (30) genes. The Tetrahymena consensus sequence has an A at -3 and a G at +4, an arrangement found in about 35% of eukaryotic messages. Interestingly, this arrangement has been found to be the most efficient initiator of translation (31) in higher eukaryotes. A search of the anticodon loop of Tetrahymena's initiator tRNA (32) and of mature 17 S Tetrahymena rRNA (33) failed to reveal any region that might pair with this consensus sequence. The sequence resembles those preceding the ATG in sea urchin histone genes (CAPyNATG), (30) and Py$[A]_{1-10}$ that precedes the ATG in Dictyostelium genes (34) and yeast histone genes (35). It seems likely that this Tetrahymena-specific consensus sequence plays some undetermined role in translation initiation of both histone and non-histone messages.

Mapping the 5' ends of Tetrahymena H4 transcripts

The extreme AT richness of Tetrahymena flanking DNA (Fig. 1) makes it difficult to limit the nuclease digestions required to map the 5' ends of transcripts using S1 nuclease (36). For this reason, we have mapped the ends of H4-I transcripts by three different techniques: S1 nuclease, RNase protection and primer extension (with three different primers). In every case, extensive heterogeneity was observed (Fig. 2). The most consistently observed major bands seen with each technique, map to the second A residue following a pyrimidine located about 50 bp upstream of the ATG. Similar studies on H4-II (data not shown) demonstrate two major start sites at the 1st or 2nd A residue following a T, 54 and 59 bp before the ATG. It is not likely that this heterogeneity is due to degradation of RNA since we have mapped discrete sites at the 3' ends of H4 messages using the RNase mapping technique (Fig. 7b). Multiple start sites at the A of TA sequences for the Tetrahymena actin messages have also been demonstrated using primer extension (Clair Cupples, PhD Thesis, York University). It seems likely, therefore, that both histone and non-histone messages in Tetrahymena begin largely at A residues in the sequence PyAA.

Multiple 5' ends of messages are not unusual either in higher or lower eukaryotes although, to our knowledge, they have not been reported previously

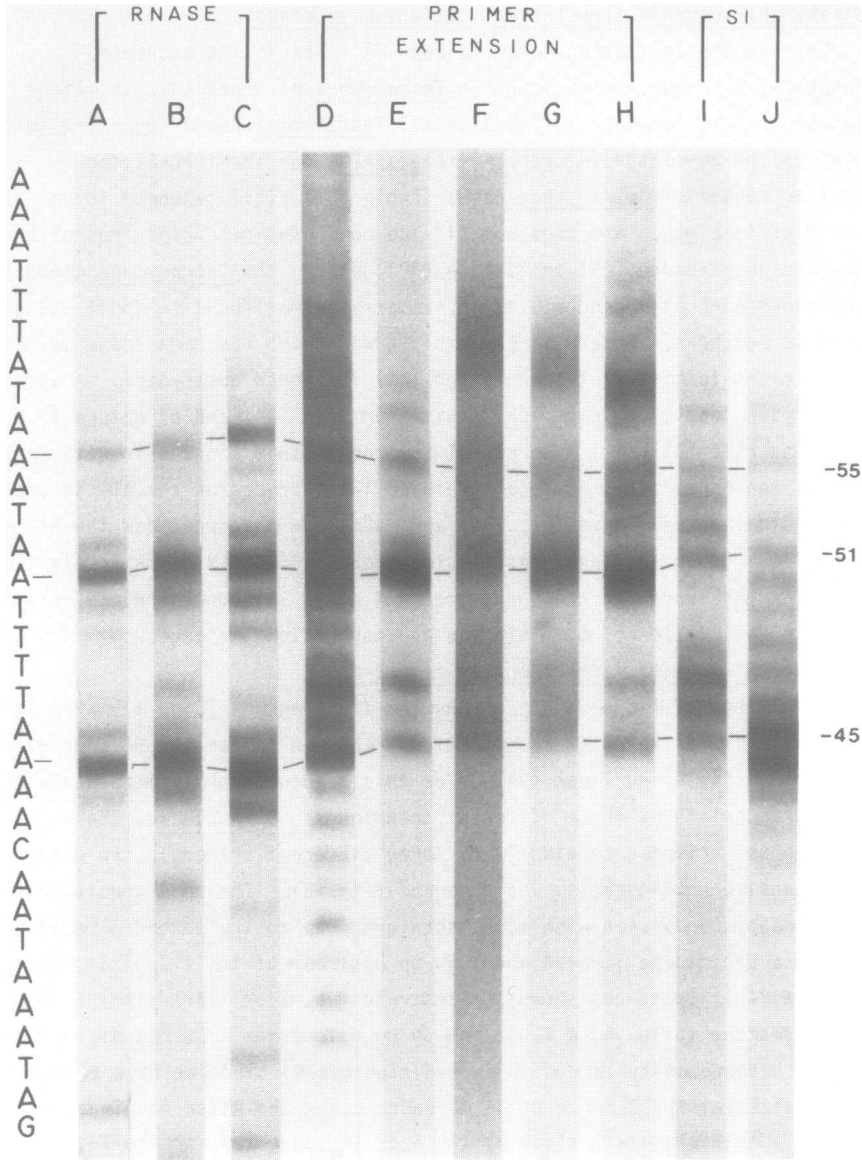

Figure 2: Transcription of H4-I Initiates at Multiple Sites. RNase Protection: ^{32}P-labeled antisense transcripts of H4-I (spanning positions -173 to +88) were made in vitro, hybridized to Tetrahymena cellular RNA and digested with RNase A and RNase T1. Lanes A + C: Growing cell cytoplasmic poly A$^+$ RNA (4 µg), digested 10 min at 0°C. Lane B: Growing cell cytoplasmic RNA (60 µg), digested for 10 min at 30°C. Primer extension: ^{32}P-end-labeled primers were annealed to cytoplasmic poly A$^+$ RNA (15 µg) from

growing cells, and extended with reverse transcriptase. (See Materials and Methods for a description of the primers). Lanes D and E: Primer 1. Duplicate experiments performed on different days. Lanes F and G: Primer 2. Duplicate experiments performed on different days. Lane H: Primer 3. S1 Nuclease Digestion: A ^{32}P-end-labeled Dde I fragment was hybridized to cytoplasmic poly A$^+$ RNA (5 µg) from growing cells, and digested for 30 min at 20°C. Lane I: 100 units/ml. Lane J: 500 units/ml. All samples were run on 6% polyacrylamide, 8.3 M urea gels, using ^{32}P-labeled DNA size markers, and visualized by autoradiography. The portion of H4-I sequence corresponding to the strongest points of initiation is shown.

for histone messages. The selection of one or more particular start sites can occur in the presence or absence of TATA boxes (37). Therefore, the existence of multiple transcription start sites does not shed any additional light on the presence or absence of these elements in Tetrahymena (see below).

The 5' ends of messages can usually be equated with the sites at which transcription is initiated, except in cases such as in Trypanosomes (38) where a leader sequence is spliced onto the ends of messages. The fact that we obtained similar results using primer extension or nuclease digestion precludes the addition of similar leader sequences on the H4-I or H4-II messages of Tetrahymena. We have also demonstrated that the 5' ends of nuclear H4-I transcripts are identical to those of cytoplasmic messages when mapped by RNase digestion (Fig. 3). These results, coupled with the similarity in sequence at the 5' ends of transcripts from H4-I, H4-II, H1 and actin genes (see below), and the fact that (aside from the aforementioned case of Trypanosomes, and capping) 5' processing of pol II transcripts has not been reported, make it likely that we have identified the starts of transcription of these genes.

Previous studies have indicated that the abundance of H4-I and H4-II mRNA's are regulated in growing and starved Tetrahymena (15). In some systems, heterogeneity of 5' ends is associated with transcriptional regulation through the use of alternative promoters (39). We have compared the 5' ends of H4-I and H4-II messages in growing and starved cells and no differences were detected (Fig. 3). We therefore conclude that the observed heterogeneity of the 5' ends of histone messages in Tetrahymena is probably a fundamental property of the transcription machinery rather than a consequence of differential regulation.

5' consensus sequences

The 5' flanking sequences of both H4-I and H4-II are extremely AT-rich (>75%). As might be expected, computer searches of this region yield numerous sequences differing by no more than one nucleotide from the TATA box consensus sequence (TATA A A A). Seven and six such matches are found 20-280 bases

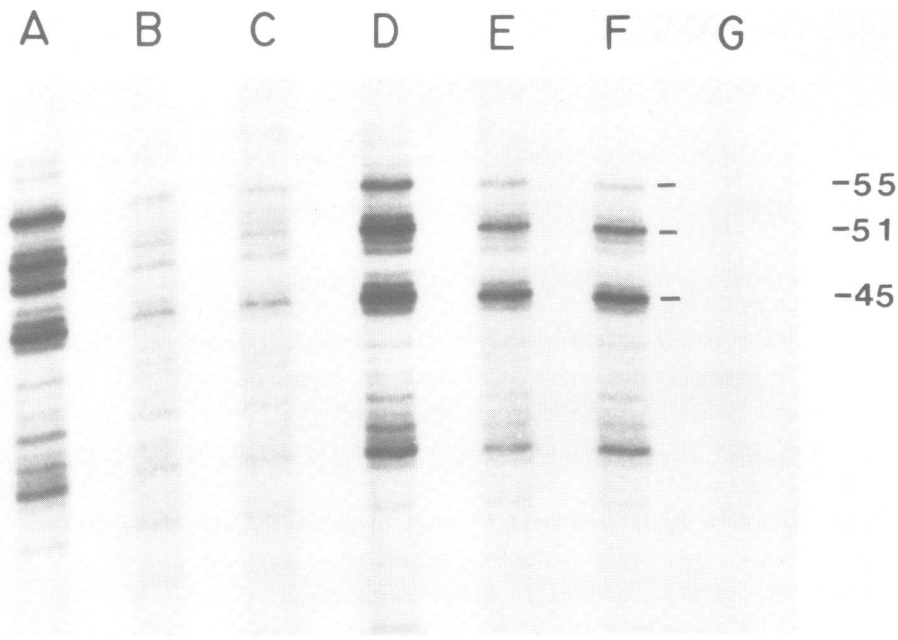

Figure 3: Heterogeneous 5' Ends are Not a Consequence of Differential Regulation.
^{32}P-labeled anti-sense transcripts of H4-I (from positions -173 to +88) were made in vitro, hybridized to RNA from different sources and digested with RNase A and T1 for 10 min at 0°C. Samples were run on polyacrylamide gels and visualized as described in Figure 2. Lane A: cytoplasmic, poly A$^+$ RNA (10 µg) from starved cells, Lane B: total, poly A$^+$ RNA (10 µg) from starved cells, Lane C: total, poly A$^+$ RNA (10 µg) from growing cells, Lane D: nuclear RNA (50 µg) from growing cells, Lane E: same as D, but 40 µg RNA, Lane F: cytoplasmic RNA (50 µg) from growing cells, Lane G: control, *E. coli* ribosomal RNA (10 µg). Major bands correspond to -45 bp, -51 bp and -55 bp from the ATG of H4-I.

upstream of the transcription start sites of H4-I and H4-II, respectively. Given the AT-richness of sequences flanking *Tetrahymena* genes and the fact that TATA sequences in lower eukaryotes can be variable distances from the start of transcription, we find it impossible to unambiguously identify TATA sequences (if they exist at all) in the H4 genes or in other ciliate genes.

Interestingly, in *Dictyostelium*, another lower eukaryote with extremely AT-rich DNA flanking its genes, putative TATA boxes are located at a relatively constant distance (60-90 bp) from the start of translation and are invariably found just upstream (2-20 bp) of a remarkable stretch (10-30 bp) of nearly pure T residues (34). No obviously similar punctuation of TATA-like sequences is found in the upstream sequences of ciliate genes analyzed to date. Without such

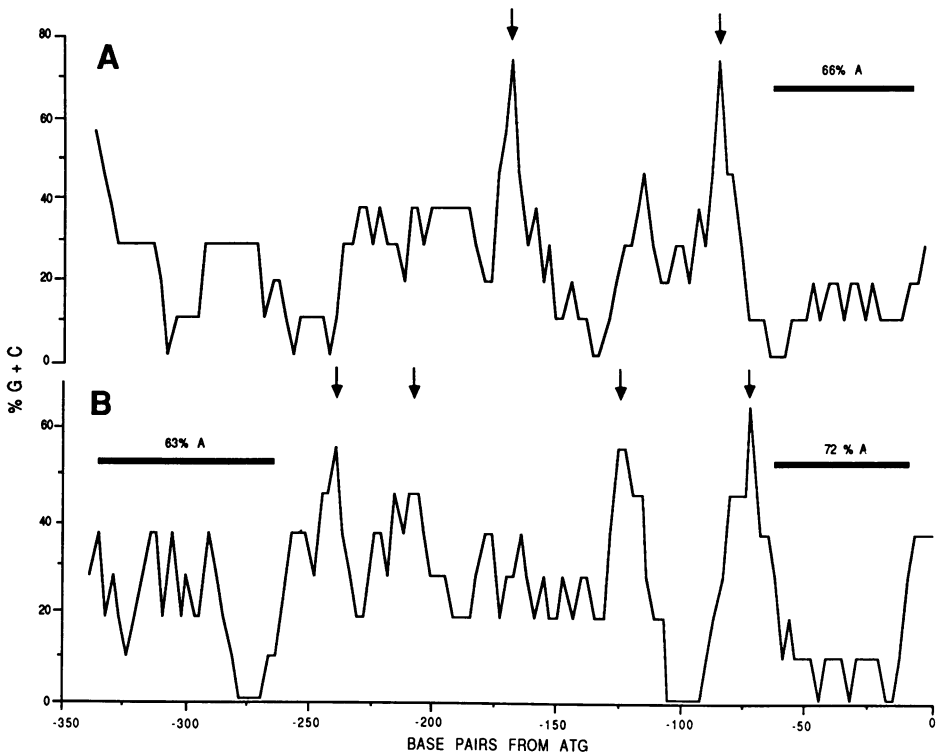

Figure 4: Base Composition Upstream of Tetrahymena H4 Genes.
The relative GC content is plotted for the coding strands in the region 340 bp upstream of H4-I (A), and 340 bp upstream of H4-II (and H3-II, although in the opposite strand) (B). Base composition was determined every third base, in a 5 bp range, using Pustell-Kafatos DNA analysis software (18). Values obtained were then re-plotted at each position using Cricket Graph, on a Macintosh personal computer. Arrows point to the peaks of GC-richness. Bars show extremely A-rich regions (in the opposite strand for H3-II). The %A was calculated by counting the A residues in the regions from -10 bp to -60 bp upstream of the H4 genes, and from -10 bp to -66 bp upstream of H3-II.

additional punctuation, it is possible that RNA polymerase II cannot distinguish among numerous, TATA-like sequences in the 5' sequence flanking Tetrahymena genes, raising the possibility that a TATA box plays little or no role in transcription initiation in Tetrahymena.

We have searched the 5' flanking sequences of H4-I and H4-II for other conserved elements of pol II genes without success; these include the "CAT" box (GG$^{C}_{T}$CAATCT); the "cap box" (CTTPyTG) and known variants of the enhancer core (GTG$^{T}_{A}$T$^{T}_{A}$G). We have also been unable to find convincing evidence for a number

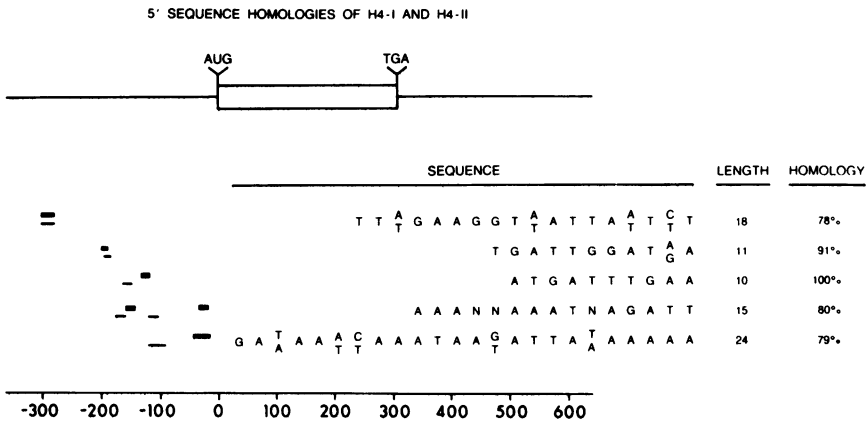

Figure 5: Homologous Sequence Elements 5' to Tetrahymena H4 Genes.
(A) 5' sequence homologies of H4-I and H4-II. The coding region is boxed. The relative positions of the sequences are noted by lines, ( ) = H4-I, ( ) = H4-II. Sequences were identified using the Autoalign program of Pustell-Kafatos software (18). Parameters of the search were as follows: range = 5, scale = .95, min. value = 60, min. length = 6, overrun = 6. Matching sequences were then scanned (by eye) to determine the best match. Short sequences containing only A + T were ignored as insignificant.

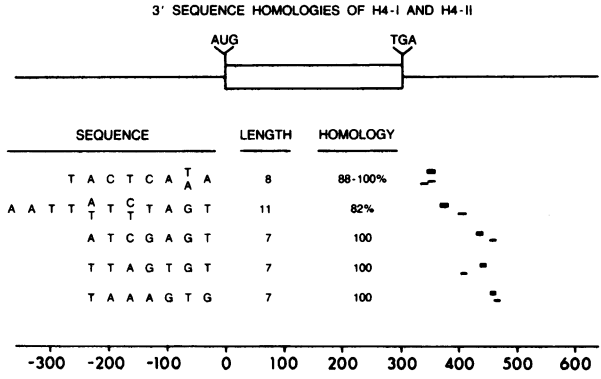

Figure 6: Homologous Sequence Elements 3' to Tetrahymena H4 Genes.
Same as Figure 5, but 3' flanking DNAs were compared.

of histone gene-specific upstream sequences including the GATCC sequence found upstream of TATA, the histone-specific "CAT" box ($G^{G}CAATN^{A}$), and the histone-specific "CAP" box ($^{P}yCATTC^{P}u$).

Small GC-rich clusters upstream of Tetrahymena genes

Given the AT-richness of the 5' regions flanking Tetrahymena genes, clusters of GC residues are statistically unlikely. However, as seen in Fig.

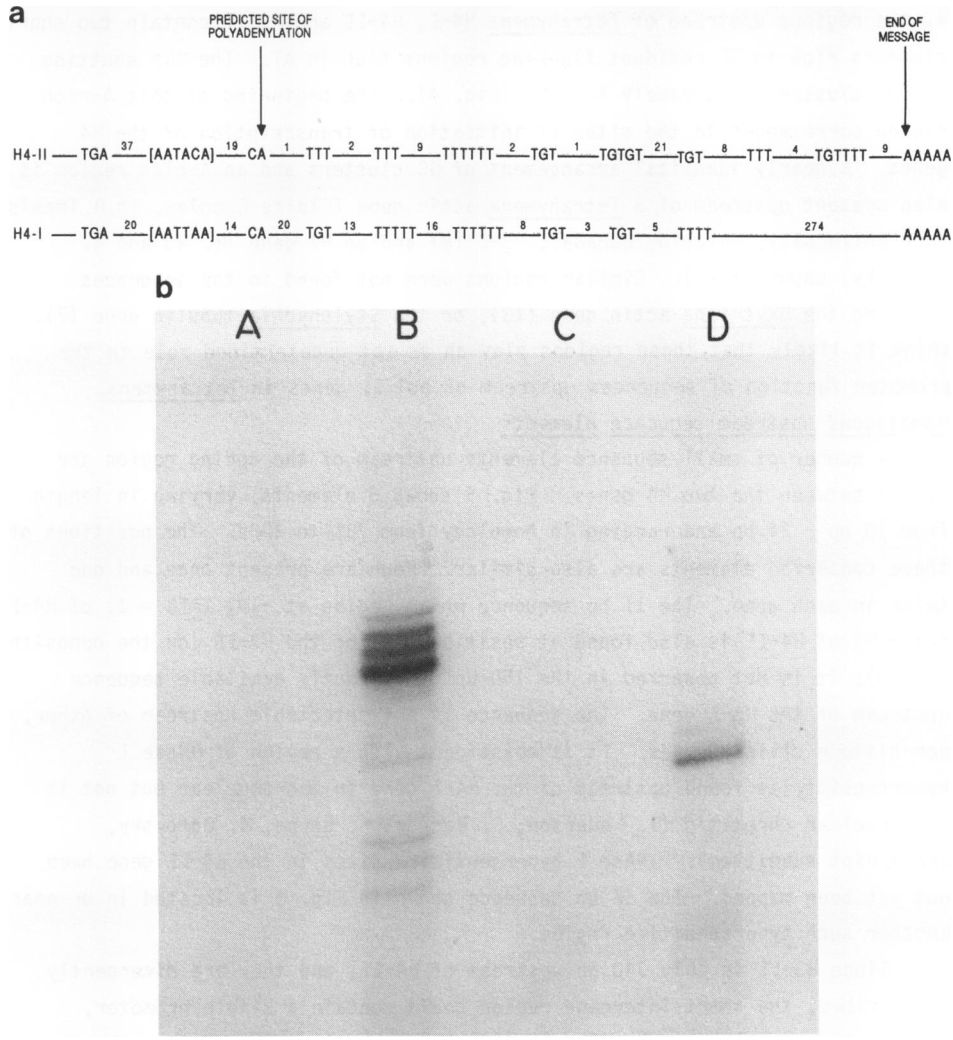

Figure 7a: A Putative Cleavage/Polyadenylation Site Downstream of Tetrahymena H4 Genes.

Sequences begin with TGA, the terminator of translation of both genes. Bracketed sequences correspond to "allowed" variations of the canonical AAUAAA (40). Arrow points to the predicted site of polyadenylation, if typical "rules" apply.

Figure 7b: 3' End Mapping of Tetrahymena H4 transcripts.

^{32}P-labeled anti-sense transcripts of H4-I (spanning positions +420 to +985) or H4-II (positions +155 to +564) were hybridized and digested with RNase (as described for Figure 3). A: Control, E. coli ribosomal RNA hybridized with H4-II transcript. B: Poly A$^+$ RNA from growing Tetrahymena cells, H4-II transcript. C: Control, E. coli ribosomal RNA, H4-I transcript. D: Same as B, but H4-I transcript. The termination site of H4-II maps to position +464. H4=I is polyadenylated at position +719.

4, the regions upstream of Tetrahymena H4-I, H4-II and H3-II contain two short clusters rich in GC residues flanking regions high in AT. The DNA abutting the 3' cluster is extremely A-rich, (Fig. 4). The beginning of this A-rich region corresponds to the sites of initiation of transcription of the H4 genes. A nearly identical arrangement of GC clusters and an A-rich region is also present upstream of a Tetrahymena actin gene (Claire Cupples, Ph.D Thesis, York University, Ontario, Canada), H3-I (8) and an H1 gene (M. Wu and M. Gorovsky, unpublished). Similar regions were not found in the sequences flanking the Oxytricha actin gene (13), or the Stylonychia tubulin gene (7). We think it likely that these regions play an as yet undetermined role in the promoter function of sequences upstream of pol II genes in Tetrahymena.

Homologous upstream sequence elements

A number of small sequence elements upstream of the coding region are shared between the two H4 genes. Fig. 5 shows 5 elements, varying in length from 10 bp - 24 bp and ranging in homology from 78% to 100%. The positions of these conserved elements are also similar. Four are present once and one twice in each gene. The 11 bp sequence which begins at -191 (ATG = 1) of H4-I and -188 of H4-II is also found at position -201 of the H3-II (on the opposite strand); it is not observed in the 160 bp of currently available sequence upstream of the H3-I gene. The sequence is not detectable upstream of other, non-histone ciliate genes. It is coincident with a region of DNase I hypersensitivity found upstream of the H4-I gene in macronuclear but not in micronuclear chromatin (D. Pederson, G. Bannon, K. Shupe, M. Gorovsky, manuscript submitted). DNAse I hypersensitive sites in the H4-II gene have not yet been mapped. The 18 bp sequence shown in Fig. 5 is located in or near another such hypersensitive region.

Since H3-II is only 340 bp upstream of H4-II, and they are divergently transcribed, the short intergene region could contain a single promoter, overlapping promoters or completely independent dual promoters in both orientations. It was therefore appropriate to compare this DNA segment with itself, its complement and its reverse complement. Aside from the GC-rich region and the 11 bp homology discussed above, this comparison has found the sequence ATCCAATCAAATTG located 147 bp upstream of the ATG of H4-II. It is also present 160 bp upstream of H3-II (in the opposite strand) but with an additional A, (ATCCAATCAAAATTG). Since these sequences are only 32 bp apart, they may interact in the regulation of these genes.

The 3' flanking regions

We have compared the downstream region of the Tetrahymena H4 genes and have

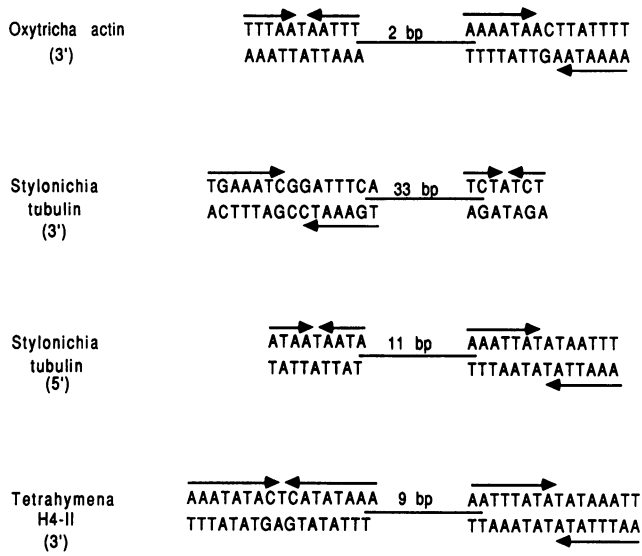

Figure 8: A Putative Origin of Replication in Ciliate Genomes.
True palindromes are noted by opposing arrows on the same strand, inverted repeats by arrows on opposite strands. 5' and 3' are relative to the coding regions. References: Oxytrichia actin (13), Stylonichia tubulin (7).

searched for termination consensus sequences (40). Fig. 6 shows 5 short (7 bp - 11 bp) sequence elements ranging in homology from 82% to 100%. As with the 5' homologous regions, these sequences are located in similar positions in the two genes. The regions flank a minor DNAse I hypersensitive site just outside the H4-I gene in macronuclear but not micronuclear chromatin. Like genes encoding other polyadenylated histone mRNAs, the Tetrahymena H4 3' flanking regions do not contain the conserved terminal hairpin structure that functions in transcript termination/processing in non-polyadenylated histone mRNAs (41). However, each H4 gene does contain a permitted variation of the canonical AAUAAA (40) followed by a short stretch of DNA ending in CA (Fig. 7a), a sequence motif frequently associated with polyadenylation (41). Downstream of this region, each gene has frequent short runs of T's interspersed with at least 3 TGT sequences. Similar sequences have been observed in genes coding for polyadenylated mRNAs in higher eukaryotes and their viruses (41), and a role for these sequences in transcription termination and/or cleavage and/or polyadenylation in Tetrahymena seems likely. In yeast, another organism whose histone mRNAs are polyadenylated, the canonical polyadenylation signal is not evident in the 3' flanking sequence, but smaller sequence elements similar to

those in Tetrahymena (TAG, TATGT, TAGT, TTT, TTTT) are scattered in the region of the approximate message 3' ends (41,42).

Mapping the 3' ends of Tetrahymena H4 Transcripts

To determine if the putative cleavage/polyadenylation signals found in the 3' regions of H4-I and H4-II are actually associated with this function in Tetrahymena in a manner similar to that in higher eukaryotes, and to provide the first data mapping the 3' end of ciliate messages, we performed RNase protection experiments using in vitro-transcribed RNA complementary to the 3' ends of the two mRNA's.

Figure 7b shows that despite the similarity in sequence motifs in the region just following the TGA of both genes, the actual sites of polyadenylation are quite different for the two messages. While the site of polyadenylation/cleavage of H4-II is relatively close to the motif (153 bp from TGA), H4-I mRNA is polyadenylated much further downstream (410 bp from TGA). Both messages are polyadenylated at sites considerably downstream of the polyadenylation sites predicted by the putative signal (following the first CA after the AAUAAA), suggesting that the mechanism by which Tetrahymena polyadenylates its messages may be different from higher eukaryotes. The 3' ends of the mRNA's do not exhibit the extensive heterogeneity seen at the 5' ends.

Secondary structure analyses

We have searched the flanking sequences of the H4-I and H4-II genes for direct repeats, true palindromes and inverted repeats. In addition to the direct repeats observed in the 5' regions of both genes (Fig. 5) the only notable structure observed is shown in Fig. 9. This structure, found in the 3' flanking sequences of the H4-II consists of a true palindrome followed by an inverted repeat (hairpin). Similar structures have been found near the ends of the macronuclear DNA molecules containing the Stylonychia α-tubulin gene (7) and the Oxytricha actin gene (13), as well as in the origin of replication of polyoma virus (43). Since the linear macronuclear DNA molecules of hypotrichous ciliates initiate DNA replication largely at one or both ends, and an ARS sequence capable of supporting autonomous replication of plasmids has been found in the 3' regions of yeast histone genes (see 5 for review), we think it likely that this structure is characteristic of at least one class of ciliate replication origins.

SUMMARY AND CONCLUSIONS

We have presented the first detailed comparison of members of a small

multigene family in ciliates. These studies have identified sequences likely to be important in transcription, translation, replication and chromatin structure in this group of ancient eukaryotes. These studies should provide essential starting points for more detailed functional analyses using DNA-mediated transformation or *in vitro* systems.

Present addresses: [1]Department of Pediatrics, University of Rochester, Rochester, NY 14627, and [2]Department of Biochemistry, University of Arkansas, Little Rock, AR 72205, USA
*To whom correspondence should be addressed

REFERENCES

1. Hayashi, T., Hayashi, H., Fusauchi, Y., and Iwai, K. (1984) J. Biol. Chem. 95, 1741-1749.
2. Hayashi, H., Nomoto, M., and Iwai, K. (1984) J. Biol. Chem. 96, 1449-1456.
3. Nomoto, M., Hayashi, H., and Iwai, K. (1982) J. Biol. Chem. 91, 897-904.
4. Fusauchi, Y. and Iwai, K. (1984) J. Biol. Chem. 95, 147-154.
5. Gorovsky, M.A. (1986) J.G. Gall, ed. Academic Press, NY. In Press.
6. Caron, F. and Meyer, E. (1985) Nature 314, 185-188.
7. Helftenbein, E. (1985) Nucleic Acids Res. 13, 415-433.
8. Horowitz, S. and Gorovsky, M.A. (1985) Proc. Natl. Acad. Sci. USA 82, 2452-2455.
9. Kuchino, Y., Hanyu, N., Tashiro, F. and Nishimura, S. (1985) Proc. Natl. Acad. Sci. USA 82, 4758-4762.
10. Preer, J.R. Jr., Preer, L.B., Rudman, B.M. and Barnett, A.J. Nature 314, 188-190.
11. Kumazaki, T., Hori, H., Osawa, S., Mita, T. and Higashinakagawa, T. (1982) Nucleic Acids Res. 10, 4409-4412.
12. Zaug, A.J. and Cech, T.R. (1986) Science 231, 470-475.
13. Kaine, B.P. and Spear, B.B. (1982) Nature 295, 430-432.
14. Bannon, G.A., Bowen, J.K., Yao, M.-C. and Gorovsky, M.A. (1984) Nucleic Acids Res. 12, 1961-1975.
15. Bannon, G.A., Calzone, F.J., Bowen, J.K., Allis, C.D., and Gorovsky, M.A. (1983) Nucleic Acids Res. 11, 3903-3917.
16. Gorovsky, M.A., Yao, M.-C., Keevert, J.B. and Pleger, G.L. (1975) Meth. in Cell Biol. 9, 311-327.
17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. 74, 5463-5467.
18. Pustell, J. and Kafatos, F.C. (1982) Nucleic Acids Res. 10, 51-59.
19. Calzone, F.J., Stathopoulos, V.A., Grass, D.G., Gorovsky, M.A. and Angerer, R.C. (1983) J. Biol. Chem. 258, 6899-6903.
20. Maniatis, T.F., Fritsch, E.F. and Sambrook, J. (1982) Cold Spring Harbor, NY.
21. Melton, D.A., Krieg, P.A., Rebagliati, M.R., Maniatis, T., Zinn, K. and Green, M.R. (1984) Nucleic Acids Res. 12.
22. Ghosh, P.K., Reddy, V.B., Piatak, M., Lebowitz, and Weissman, S.M. (1980) Meth. in Enzymol. 65, 580-595.
23. Benyajati, C. and Dray, J.F. (1984) Proc. Natl. Acad. Sci. USA. 81, 1701-1705.
24. Eickbush, T.H. and Burke, W.D. (1986) J. Mol. Biol., In press.
25. Doolittle, W.F. (1985) Trends in Genetics. 1, 64-65.
26. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
27. Ikemura, T. and Ozeki (1983) Cold Spring Harbor Symp. Quant. Bio. 47, 1087-1097.

28. Nussinov, R. (1982) Biochem. Biophys. Acta 698, 111-115.
29. Kozak, M. (1984) Nucleic Acids Res. 12, 857-872.
30. Hentschel, C.C. and Birnstiel, M.L. (1981) Cell 25, 301-313.
31. Kozak, M. (1986) Cell 44, 283-292.
32. Kuchino, Y., Mita, T., and Nishimura, S. (1981) Nucleic Acids Res. 9, 4557-4562.
33. Engberg, J., Din, N., Saiga, H., and Higashinakagawa, T. (1984) Nucleic Acids Res. 12, 959-972.
34. Barklis, E., Pontius, B., Barfield, K. and Lodish, H.F. (1985) Mol. and Cell. Biol. 5, 1465-1472.
35. Smith, M.M. and Andressen, O. (1983). J. Mol. Biol. 169, 663-690.
36. Berk, A.J. and Sharp, P.A. (1977) Cell 12, 721-732.
37. Nagawa, F. and Fink, G.R. (1985) Proc. Natl. Acad. Sci. USA. 82, 8557-8561.
38. Milhausen, M., Nelson, R.G., Sather, S., Selkirk, M. and Agabian, N. (1984) Cell. 38, 721-729.
39. Clerc, R.C., Bucher, P., Strub, K., and Birnstiel, M.L. (1983) Nucleic Acids Res. 11, 8641-8657.
40. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
41. Birnstiel, M.L., Busslinger, M. and Strub, K. (1985) Cell 41, 349-359.
42. Henikoff, S., Kelly, J.D. and Cohen, E.H. (1983). Cell 33, 607-614.
43. Soeda, E., Miura, K., Nakaso, A. and Kimura, G. (1977) FEBS Lett. 79, 383-389.
44. Madura, K. and Prakash, S. (1986) J. Bact. 166, 914-923.
45. Reynolds, P. Higgins, D.R., Prakash, L. and Prakash, S. (1985) Nucleic Acids Res. 13, 2357-2372.